DOCUMENT RESUME

ED 021 464 24 EM 000 284

By- Gorth, W. P.; And Others

VALIDATION OF A CRITERION OF LECTURE EFFECTIVENESS. RESEARCH MEMORANDUM.

Stanford Univ., Calif. Stanford Center for Research and Development in Teaching.
Spons Agency- Office of Education (DHEW), Washington, D.C. Bureau of Research.
Report No- SU-SCRDT-RM-26
Bureau No- BR-5-0252
Pub Date Mar 68
Contract- OEC-6-10-078
Note- 29p.
EDRS Price MF-$0.25 HC-$1.24
Descriptors- ACADEMIC ABILITY, ACHIEVEMENT, *LECTURE, *RESEARCH METHODOLOGY, TEACHING SKILLS, TEACHING TECHNIQUES, *VIDEO TAPE RECORDINGS

Studies of achievement scores as the criterion of lecture effectiveness have been limited to use of experienced teachers lecturing in the classroom to classes of one age group only. This study sought (1) to compare videotape recordings of lectures with live lectures, (2) to determine whether the quality of videotape affects achievement scores, and (3) to investigate the interaction effects of test validity, students' ability and age, and number of presentations of the lecture. Lecture effectiveness was defined as the ability to explicate ideas to students so that they are able to answer questions about these ideas. The complete factorial design initially included 20 groups of about 20 students each. Analysis of covariance of criterion test scores showed that high scores were correlated with the viewing of videotapes of high quality and with the viewing of effective lectures by high ability students. Repetition of the same lecture on videotape intensified the variations in lecture effectiveness. It was concluded that his experiment offers a partial validation for the use of videotape to represent live classroom lectures in research on lecture effectiveness. (LH)

STANFORD CENTER FOR RESEARCH AND DEVELOPMENT
IN TEACHING

Research Memorandum No. 26

VALIDATION OF A CRITERION OF LECTURE EFFECTIVENESS

by

W. P. Gorth
Stanford University

D. W. Allen
University of Massachusetts

L. W. Popejoy and T. W. Stroud
Stanford University

This memorandum is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The memorandum should not be cited as a reference without specific permission of the authors. It is automatically superseded upon formal publication of the material. The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education, under the provisions of the Cooperative Research Program

School of Education
Stanford University
Stanford, California

March 1968

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED021464

EM 000 284

PREFACE

This paper deals with several problems of research on the teacher's lecture effectiveness. In previous studies the criterion of effectiveness has been a measure of student achievement administered after presentation of live lectures. Validation of this criterion has been attempted through postdictions of mean scores on these measures from ratings of videotapes of the live lectures. The present paper discusses certain problems inherent in this procedure and investigates alternative methods.

Specifically, the paper considers two basic questions: (1) whether live and videotape recordings of lectures produce the same relative mean achievement scores and (2) whether the quality of videotape recordings affects student's scores on the criterion measure. Several additional questions concerned with effects of student's ability, age of students, and the nature of the criterion itself are also investigated.

In view of the questions with which the paper deals, it seems to have relevance for an area of research within the program of the Center. For this reason it is being distributed as a Research Memorandum. Since there may be problems of clarity and consistency in the present version, the author's solicit critical comments and suggestions for improvement prior to publication in an appropriate journal.

Richard Lindeman

VALIDATION OF A CRITERION OF LECTURE EFFECTIVENESS[1]

In this study the lecture effectiveness of teachers is defined as the ability to explicate ideas to students so that they are able to answer questions about these ideas. This study and previous studies of the lecture effectiveness of teachers (Fortune, Gage and Shutes, 1966; Podlogar, Rosenshine, and Gage, 1967; Unruh, 1967) have followed the criterion of effectiveness paradigm, In this paradigm, the independent variables are (1) prediction sources, or all teacher variables existing before the classroom performance, (2) contingency factors, or subject matter, environmental and pupil variables, and (3) classroom behaviors of teachers and students. The dependent variable, or criterion of effectiveness, is change in student achievement of intermediate goals of education. In the previous studies cited above the predictive source has been limited to experienced teachers. The contingency factor has been limited to one age level in either randomly assigned groups or in high school classes and to narrowly defined content. The classroom behavior of the teacher has been limited to lecturing. The criterion of effectiveness has been the mean score on a ten-item achievement test earned by students hearing the lecture.

The basic questions considered in this paper are (a) whether live and videotape recordings of lectures produce the same relative mean achievement scores and (b) whether the quality of a videotape recording affects the students' scores. Although videotape recordings make it possible to rate and categorize the same classroom behavior repeatedly, they are only an abstraction of real life and have certain restrictions. Hence, the effects of recordings should be compared with those of live lectures (Gorth and Baker, 1967).

The motivation for the validation is to reject plausible rival hypotheses which confounded earlier work. If only post dictions of the criterion measures are made to validate the use of videotape recordings in research on lecture effectiveness, it is possible that the raters may be provided with information which is correlated with the lecturer's effectiveness, but does not directly contribute to it. For example, in Unruh's study (1967), the videotapes presented views of the class to which the teacher was lecturing and also varied in recording quality. Students' achievement is correlated with self-reported attention during a lecture (Fortune, Gage, and Shutes, 1961), and students' self-reported attention is correlated with their observed attending behavior (MacGraw, 1965). Thus, presenting the raters with views of the class would permit their post dictions of test scores in part from their awareness of the class's attending behavior. Similarly, the quality of the videotape is positively correlated with the lecturer's effectiveness; thus the accuracy of ratings may be in part, a function of videotape quality.

The paper considers several additional problems. One is to determine whether the results of previous research, in which subjects were high school seniors hold also when subjects are eighth and ninth grade students. In this experiment the prediction sources and classroom behaviors were identical with those in the study by Podlogar, Rosenshine, and Gage, (1967), (because videotapes of the live lectures were used.) However, the contingency factor, subjects' age, was modified. The effect of this modification was judged in terms of the degree to which the lecturers' ranks remained the same as in the previous studies.

The effect of varying the verbal and quantitative abilities of students was also investigated. The question was whether student ability interacts with lecture effectiveness and videotape quality. If not, then the influence on achievement of lecture effectiveness and videotape quality could be considered constant over a wide range of student ability.

A further problem with which this paper deals is that of the validity of the ten item achievement test used as the criterion of effectiveness in the study of Podlogar, Rosenshine, and Gage (1967). Teachers who concentrated on fewer topics may have provided more information on them than other teachers who covered more topics but covered each less thoroughly. Conceivably, the effects of such differences could be controlled by making adjustments in the test based on the content of the lectures; however, then the criterion measures would no longer be equivalent and the comparisons among lectures would be ambiguous.

To deal with this problem test items were constructed to measure achievement in each section of the article on Yugoslavia. These detailed measures provided information with which to decide if lecture effectiveness varied across closely related topics. The degree of variation provided a measure of the specificity of effectiveness to topics. Results on these measures were compared with those on the original ten item criterion.

A final question investigated was whether scores on the criterion measures were a function of the number of presentations of the lecture. In the usual live classroom situation a lecture is presented and then followed by a criterion test.

When videotape is used, other possible schedules of presentation and testing may be found to yield more sensitive measures of lecture effectiveness. For example, lectures may be presented repeatedly until mean achievement reaches a maximum. Lecture effectiveness could then be defined as a function of both achievement and number of presentations. This experiment investigated achievement after one and after two viewings of a videotape and considered lecture effectiveness as a function of the number of presentations as well as of the level of achievement reached.

METHOD

Design

In this study, each of 20 groups of about 20 students, 10 groups in the morning and 10 in the afternoon of the same day, wrote a 10 minute pre-test, viewed a 15-minute videotaped lecture, wrote a 10-minute mid-test, viewed a 15-minute videotaped lecture, and wrote a final 10-minute post-test.

Each of four videotaped lectures was distributed among twenty groups of students during the first tape presentation, and distributed again during the second presentation, in such a way that each of the tapes was seen by five classes at each presentation, and that every one of the sixteen possible ordered samples of two tapes (drawn with replacement) was presented. Four of the combinations were chosen for duplication and presented to the four remaining groups.

One booklet containing three tests was distributed to each student. The booklets forced the students to take the tests in a particular order. The format of the tests was similar, so that different tests could be given to different

students at the same time without their knowledge. Two parallel forms of the criterion test and one science achievement test were used, so that another experiment could be performed simultaneously which considered the relation of repeated administration of relevant achievement tests (Gorth, Allen, Popejoy, and Stroud, 1963). Eighteen different booklets were collated. They included all of the possible combinations of a pre-test (either the science test or one of the two parallel forms of the criterion test), a mid-test (again either science or the criterion tests), and a post-test (one of the two forms of the criterion test). To insure that they would be distributed randomly to the students in the classrooms, a different, randomly arranged stack of the 18 booklets plus a random selection of extra booklets at the bottom of the stack was distributed randomly to each group of students.

Setting

Ten rooms of an elementary school were fitted with television monitors and videotape recorders. Five of these ten classrooms contained a 35 mm time lapse camera which photographed the students during the entire experiment. Each room was used once in the morning and once in the afternoon.

Subjects

Initially about 400 subjects were hired from the local schools from the eighth and ninth grades. They were paid to participate in the Stanford Teacher Education Program for 15 hours distributed evenly over three consecutive days during the summer of 1967. About 200 students asked specifically to work in the morning, and the rest preferred to work in the afternoon. The morning

and afternoon students were each divided randomly into 10 groups of about 20 students which were assigned to rooms. Each student had taken the Necessary Arithmetic Operations, R-4, and the Wide Range Vocabulary Test, V-3 (French, Ekstrom, and Price, 1963) during the first hour of the first day of their job. They were taught by teaching interns for the next 13 hours on the three days. The experiment was carried out during the last hour of the third day. The proctors who supervised the testing were graduate students in education.

Videotape Selection

The four selected videotapes of teachers' lectures on Yugoslavia included two which portrayed less effective lectures (25Y and 31Y) and two which portrayed more effective lecutres (13Y and 18Y). The ranking of effectiveness was based upon the mean scores of the students who viewed the live lectures and was provided by Podlogar (1967).

When choosing the videotapes, a large variation was found in the quality of the recording. Some of the tapes showed a one-half or a one-quarter screen picture of the teachers or had "snow". The quality of the videotapes was considered good if the sound and picture were clear and considered poor if there were "snow" and distortion. Quality was ranked by four experimenters independently. The following four tapes ranked in quality from best to worst respectively, 31Y, 18Y, 25Y, and 13Y (31Y and 18Y appreciably better than 25Y and 13Y). The two lectures to be shown in each room were copied onto one tape.

Achievement Tests

The measure of effectiveness of the teachers' lectures was the mean student achievement on two parallel forms of a thirty-item, four-alternative, multiple-choice achievement test, the criterion test, given during the 10 minutes immediately following the presentation. The first ten items on one test form were identical with those used by Fortune, Gage, and Shutes (1966). Ten similarly worded items were written by the experimenters and distributed randomly among the items on the second form. The twenty additional items needed to complete each test were selected randomly from a pool of forty items written by the experimenters[2] and arranged randomly on each test. The forty items had been selected from seventy items which were pretested on 47 high school students not involved in the experiment. The material measured by the questions was uniformly distributed throughout the article on Yugoslavia from which the teachers had prepared their lectures. The article contained five selections of content denoted by headings which served as criteria for grouping items into scales to measure achievement in each section.

RESULTS

Analysis of the Pre-Test

In all analyses of covariance described below, the student's score on the Wide Range Vocabulary Test, the verbal aptitude Test, and The Necessary Arithmetic Operations Test, the quantitative aptitude Test, are entered as covariates. The dependent variable is the total number of items on a criterion

test answered correctly. Some students took science tests for the pre-test and the mid-test. Their scores on these tests are not included in the analysis.

In the pre-test analysis, the possible effects on pre-test means of three discrete variables are considered. The first was the room in which the presentations and tests were taken. This was included to detect variance which may be assigned to the effects of different proctors or to the cohesiveness the group may have developed after 14 hours together. The second was the time of day; the purpose was to detect variance caused by the self-selection of different students for the morning or the afternoon. The third variable was the form of the criterion test, which was included in order to detect variance due to unequivalent forms.

The results of the analysis are shown in Tables 1, 2, and 3.

Insert Tables 1, 2, 3 about here

The partial correlation of the verbal aptitude test with the criterion test, holding the quantitative score constant, is significant at the .05 level. The partial correlation of the quantitative test with the criterion test, holding the verbal aptitude test constant, is significant at the .001 level. The discrete variables are all nonsignificant. The mean total score on the criterion tests for students is 8.44 after adjustment by the analysis of covariance.

Analysis of the Mid-Test

The dependent variables, the mean score of the students on the 30-item criterion test and the 10-item test of Fortune, Gage, and Shutes (1966) were

analyzed across lectures. Videotape playback problems prevented the showing of the videotapes in one room and the administration of the mid-tests or the post-tests in one room. As shown in Table 4, the analysis of both variables showed us significant differences.

Insert Table 4 about here

Analysis of the Post-Test

A survey of the class-by-class multiple correlation coefficients between the post-test score and the verbal and quantitative aptitude scores revealed, while the highest 16 multiple correlations were clusterd between .5292 and .7737, the lowest three had the values .4248, .3294, and .2661, producing a more markedly negatively skewness to the distribution than would be normally expected. While no external evidence was available to cast doubt on the validity of the scores form the class with the .4243 correlation, the 35 mm time-lapse photographs obtained from the class with the .3294 correlations showed that the students' attending behavior during the second videotape was very poor; they were apparently distracted by activity that was taking place outside the classroom. For this reason this class was eliminated from the post-test analysis. Although no information was available on the class with the .2661 correlation, this class was also rejected because of the very low correlation, leaving 17 classes in the study. The results are shown in Table 4. When the post-test scores are grouped by the lecture the students viewed during the first presentation, there is no significant difference in lecturers. When the post-test scores are grouped by the lecture the students viewed during

the second presentation, there is signficant difference at the . 025 level. When the two more effective lectures and the two less effective lectures are grouped together, the dependent variable is not significantly higher for the more effective lecturers, but the differences in mean scores is in the proper direction, as shown in Table 5.

Insert Table 5 about here

When the two good quality tapes and the two poor quality videotapes are grouped together, the dependent variable is significantly higher for the two good quality videotapes at the . 02 level as shown in Table 6.

Insert Table 6 about here

Comparison of Good Students and Poor Students

The students were ranked according to their scores on the best linear predictor of the criterion test score based on verbal and quantitative aptitude scores. This predictor as computed by an analysis of covariance is $Y_p = 3.87 + 0.21X_1 + 0.52X_2$, where X_1 = verbal score and X_2 = quantitative score. The top one-fourth of the students, based on Y_p, are referred to below as high ability students; the bottom one-fourth are referred to as low ability students.

Separate analyses of covariance are performed on the high and low ability students, as to the effects of the first lecturer and of the second lecturer for 17 rooms as shown in Table 7.

Insert Table 7 about here

Neither the first nor the second lecturer appears signficant in this analysis.

In order to look at the cumulative effect of two lecturers, an analysis considering the student's exposure to at least one effective lecturer effectiveness as judged by the live presentations, or to at least one videotape of good quality is presented in Tables 8 and 9 respectively.

Insert Tables 8 and 9 about here

For the low ability students, exposure to at least one good quality videotape versus no good quality tape is significant at the .025 level, while exposure to at least one effective lecturer is not significant. For the high ability students, the effect on the dependent variable of exposure to at least one good quality videotape versus no good quality tape is significant at the .002 level, while exposure to at least one effective lecturer versus none is significant at the .05 level.

Analysis by Content

Means and standard deviations are produced, by lecturer, for the score on the items based on each of the five content sections of the Atlantic article. For comparability purposes, all scores were normalized to 30 points. Table 10 shows the breakdown of scores of the test after the first presentation, tabulated by first lecture viewed and Table 11 shows the breakdown of scores of the test

after the second presentation tabulated by second lecture viewed. The tables indicate that the teachers' coverages of the five sections was uniform.

Insert Tables 10 and 11 about here

Discussion

The use of videotape recordings as representations of live classroom lecture behavior received a partial validation in our experiment. The two more effective lecturers combined received a higher mean student score than the two less effective lectures on both the mid-test and the post-test, although the differences were not significant. The rank order of the lecturers' effectiveness had changed, but these reversals can be plausibly explained by the effects of the videotape quality. It should be noted that these differences in mean scores were obtained even though the variability in achievement scores was markedly lower than that of the seniors used in the previous studies. This smaller variability would tend to make differences more difficult to detect.

The technical quality of the videotape recordings affected the performance of the students on the mid-test and the post-test. The mid-test scores of the students who has viewed a good quality tape were higher than those who had seen the poor quality tapes, although the difference was not significant. The post-test difference between the groups was significant at the .02 level. Variations in quality of the videotapes have a very large effect which cannot be ignored in research and may easily mask significant differences in the information recorded on the tapes.

The possibility of a differential effect of the lectures and of the videotape quality on students of high and low ability was considered. No significant difference was found on the post-test across lecturers for either independent variable. To decide whether the students' history of videotape viewing had masked differences in achievement, a further comparison was made for the groups of high ability students between the groups which had seen either none or at least one videotape of a successful teacher. The high ability group which viewed at least one effective lecturer achieved a significantly higher mean score than that achieved by the group who viewed no effective lectures, while no significant difference between mean scores was observed for comparable groups of low ability students. Videotape lectures which present a difficult topic in social studies to students who are younger than those in the live presentation, are differentiated in effectiveness more by the achievement of the high ability students than low ability students. The more effective lectures do indeed achieve higher mean student scores.

For high and low ability students viewing good and poor quality videotapes both ability groups achieved significantly higher mean scores if they viewed at least one good quality videotape than if they viewed none. The greater effect was seen in the scores of the high ability students, as measured by their higher level of significance of the difference in the mean scores, than for low ability students.

The criteria of effectiveness included the mean achievement score on a large set of items which were of the factual type, on items of sections of the material covered in the lecture, and on items for each section and for the entire set of items after more than one viewing of a videotaped lecture. If a comparison of mean scores is made between the pair of more effective lecturers and the pair of less

effective lecturers for each section, the pair of more effective lecturers have a higher mean score for sections one through four on the mid-test, but only one through three on the post-test. A change in relative effectiveness in specific sections is apparent. From the mean, adjusted pre-test score for all students of 8. 44 the students viewing only the two more effective lecturers had an adjusted mean score on the 30-item tests of 13. 33 for the mid-test and of 15. 33 on the post-test while the students viewing only the two less effective lecturers had 12. 55 for the mid-test and 13. 28 for the post-test. Even after two viewings the students, who only had the information provided by the less effective lecturers, did not average as high a score as the students who had viewed once the effective lecturers. The repetition of the same lecture by videotape playback intensifies the contrast between more and less effective lecturers, thus providing a more sensitive measure of effectiveness. Presumably, after several repetitions of the tape each group would reach an upper limit in their achievement scores, which might be the most stable measure of effectiveness.

In conclusion, our experiment offers a partial validation for the use of videotape recordings of teachers' lectures for research in lecture effectiveness. The variations in recording quality of the videotapes have a striking affect on students' achievement and to a different degree for high and low ability students.

References

Fortune, J.C., Gage, N.L., and Shutes, R.E. Generality of the ability to explain. Paper presented at the meeting of the American Educational Research Association, Chicago, Illinois, February 1966.

French, J.S., Ekstrom, R.B., and Price, L.A. Manual for kit of reference tests cognitive factors. Princeton, New Jersey: Educational Testing Service, 1963.

Gorth, W.P. and Baker, K. Extensions of interaction analysis of classroom behavior. In W.P. Gorth and G. Salomon (Compilers), Social psychology of education. Stanford University: School of Education (Ditto), 1967.

Gorth, W.P., Allen, D.W., Popejoy, L.W., and Stroud, T.W. The relation of repeated administration of achievement tests on students' achievement, in preparation.

MacGraw, F.W. The use of 35 mm. time-lapse photography as a feedback and observation instrument in teacher education. Unpublished doctoral dissertation, Stanford University, 1965.

Podlogar, M. Private Communication, 1967.

Podlogar, M., Rosenshine, B., and Gage, N.L. The teacher's effectiveness in explaining: evidence on its generality and correlation with students' ratings. Research Memorandum No. 10, Stanford, California: Stanford Center for Research and Development in Teaching, 1967 (mimeographed).

Unruh, W.R. The modality and validity of cues to lecture effectiveness. Unpublished doctoral dissertation, Stanford University, 1967.

FOOTNOTES

1 The authors are indebted to Dr. N. L. Gage, Dr. Richard Lindeman and Mr. Gavriel Salomon for valuable editorial assistance.

2 The items were pr epared by William Phillip Gorth and Lee W. Popejoy.

TABLE 1

Analysis of Pre-test by Room

	Room Number									
Statistic[a]	2	6	7	8	9	10	11	12	13	14
N	26	25	27	25	25	28	25	28	26	24
$\overline{X}_1$	17.15	16.56	16.52	17.28	17.40	13.21	13.92	16.21	16.04	16.54
$\overline{X}_2$	13.65	13.36	13.30	13.92	14.76	13.35	14.08	14.36	14.27	14.00
$\overline{Y}_r$	8.58	8.40	8.22	9.2	8.72	7.96	9.08	8.36	8.15	7.67
S_y	2.25	3.7	3.65	3.35	3.22	2.44	4.22	3.91	4.88	3.58
$\overline{Y}_a$	8.60	8.51	8.34	9.18	8.31	8.35	9.24	8.18	7.94	7.63

Note: The adjusted F ratio for rooms is 0.599 which is not significant.

a_N is the sample size.

$\overline{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3.

$\overline{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.

$\overline{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\overline{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for $\overline{X}_1$ and $\overline{X}_2$

TABLE 2

Analysis of Pre-test by Time of Day

	Time	
Statistic[a]	Morning	Afternoon
N	126	133
$\bar{X}_1$	14.91	17.14
$\bar{X}_2$	13.67	14.05
$\bar{Y}_r$	8.48	8.38
S_y	3.58	3.16
$\bar{Y}_a$	8.61	8.25

Note: The adjusted F ratio for time of day is 0.763 which is not significant.

a_N is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3.

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for $\bar{X}_1$ and $\bar{X}_2$.

TABLE 3

Analysis of Pre-test by Form

Statistic[a]	Form 1	Form 2
N	134	125
$\bar{X}_1$	16.07	16.04
$\bar{X}_2$	13.54	14.21
$\bar{Y}_r$	8.36	8.50
S_y	3.15	3.60
$\bar{Y}_a$	8.44	8.42

Note: The adjusted F ratio for test forms is 0.001 which is not significant.

[a] N is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3.

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, N-4.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for $\bar{X}_1$ and $\bar{X}_2$.

TABLE 4

Analysis of Mid-Test and Post-Test by Lecturer

Lecturer	Statistic[a]	Mid-test (19 rooms) 10-item test[b]	Mid-test (19 rooms) 10-item test[c]	Mid-test (19 rooms) 30-item tests	Post-test (19 rooms) Lecturer first showing	Post-test (19 rooms) Lecturer second showing
13Y	N	13	30	59	100	103
	$\bar{Y}_r$	8.14	5.43	13.07	14.09	14.56
	S_y		2.64	5.57	5.29	5.35
	$\bar{Y}_a$	.6977[d]	5.31	12.96	14.12	14.14
15Y	N	14	24	48	80	62
	$\bar{Y}_r$	8.47	5.92	14.17	14.75	14.39
	S_y		2.92	4.68	4.78	4.38
	$\bar{Y}_a$	.6264[d]	5.77	13.79	14.35	14.83
25Y	N	23	28	58	78	98
	$\bar{Y}_r$	6.54	4.36	11.81	14.58	13.38
	S_y		2.50	5.41	5.04	4.86
	$\bar{Y}_a$	-.1431[d]	4.42	11.80	14.54	13.07
31Y	N	19	28	59	85	80
	$\bar{Y}_r$	6.13	5.04	13.00	13.35	14.48
	S_y		1.77	4.16	4.75	5.08
	$\bar{Y}_a$	.4308[d]	5.14	13.29	13.71	14.68
	F_a		1.95	1.98	0.70	3.22
	$p <$		.15	.15	NS	.025

[a] N is the sample size.
$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3
$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.
$\bar{Y}_r$ is the mean number of items correct on the criterion tests.
S_y is the standard deviation of the number of items correct on the criterion tests.
$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for $\bar{X}_1$ and $\bar{X}_2$
F_a is the F ratio adjusted by an analysis of covariance.

[d] Deviation from the grand mean of raw scores adjusted for student ability. Obtained from Mrs. Maria Podlogar.

[a] Means the 10-item test included as first items of criterion test.

The 10-item criterion test was administered immediately after the recording of the videotapes (Podlogar, Rosenshine and Gage, 1966).

TABLE 5

Analysis of Mid-test and Post-test by Lecturer Effectiveness

Item	Statistic[a]	Mid-test (19 rooms)	Post-test (17 rooms)	
			Lecturers, first showing	Lecturers second showing
Effective	N	107	180	165
Lecturer	$\bar{Y}_r$	13.56	14.38	14.50
(13Y & 18Y)	S_y	5.19	5.06	5.00
	$\bar{Y}_a$	13.38	14.24	14.49
Ineffective	N	117	163	178
Lecturer	$\bar{Y}_r$	12.41	13.94	13.87
(31Y & 25Y)	S_y	4.84	4.91	4.98
	$\bar{Y}_a$	12.54	14.12	13.87
	F_a	2.01	0.07	2.04
	$p <$	.20	NS	.20

[a] N is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3.

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for $\bar{X}_1$ and $\bar{X}_2$.

F_a is the F ratio after adjustment by an analysis of covariance.

TABLE 6

Analysis of Mid-test and Post-test by Video-tape Quality and Order of Presentation

Item	Statistic[a]	Mid-test (19 rooms)	Post-test (17 rooms)	
			Video-tape, first showing	Video-tape second showing
Good	N	107	165	142
Video-tape	$\bar{Y}_r$	13.52	14.03	14.44
Quality	S_y	4.42	4.80	4.78
(18Y & 31Y)	$\bar{Y}$	13.54	14.03	14.75
Poor	N	117	178	201
Video-tape	$\bar{Y}_r$	12.44	14.30	13.99
Quality	S_y	5.51	5.17	5.14
(13Y & 25Y)	$\bar{Y}_a$	12.38	14.33	13.51
	F_a	3.63	0.48	6.59
	$p <$	.10	NS	.02

[a] N is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for $\bar{X}_1$ and $\bar{X}_2$.

F_a is the F ratio adjusted by an analysis of covariance.

TABLE 7

Analysis of Post-test by Student Ability and Teacher

Lecturer	Statistic[a]	Video-tape first showing		Video-tape second showing	
		High ability students	Low ability students	High ability students	Low ability students
13Y	N	24	25	30	28
	$\bar{Y}_r$	18.08	10.04	18.87	9.89
	S_y	3.82	4.63	3.44	3.10
	$\bar{Y}_a$	18.24	10.26	18.76	9.79
18Y	N	21	21	16	14
	$\bar{Y}_r$	19.14	12.00	18.50	12.14
	S_y	3.52	3.21	3.24	3.80
	$\bar{Y}_a$	18.99	11.83	18.73	11.66
31Y	N	18	28	15	25
	$\bar{Y}_r$	18.44	10.86	19.13	11.00
	S_y	3.28	3.93	3.40	4.04
	$\bar{Y}_a$	18.52	10.57	18.54	11.32
25Y	N	22	18	24	25
	$\bar{Y}_r$	17.86	9.39	17.17	10.16
	S_y	3.41	3.16	3.71	4.52
	$\bar{Y}_a$	17.85	9.78	17.57	9.67
	F_a	.04	1.02	0.70	1.34
	$p <$	NS	NS	NS	NS

[a] $_N$ is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of coariance fo $\bar{X}_1$ and $\bar{X}_2$.

F_a is the F ratio adjusted by an analysis of covariance.

TABLE 8

Analysis of Post-test by Student Ability and Lecturer Effectiveness

Item	Statistic[a]	High ability Students	Low ability Students
Viewed at	N	66	70
least one	$\bar{Y}_r$	18.53	10.80
effective	S_y	3.59	3.85
lecturer	$\bar{Y}_a$	19.67	11.10
Viewed	N	18	22
no	$\bar{Y}_r$	17.94	10.00
effective	S_y	3.24	4.08
lecturer	$\bar{Y}_a$	17.14	10.12
	F_a	5.75	0.88
	$p <$	.05	NS

[a] N is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the standard deviation of the number of items correct on the criterion tests.

$\bar{Y}_a$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for X_1 and X_2.

F_a is the F ratio adjusted by an analysis of covariance.

TABLE 9

Analysis of Post-test by Student Ability and Video-tape Quality

Item	Statistic[a]	High ability Students	Low ability Students
Viewed at	N	50	63
least one	$\bar{Y}_r$	19.12	11.35
good quality	$S_{\bar{y}}$	3.29	3.80
video-tape	Y_a	20.12	11.97
Viewed	N	34	29
no	$\bar{Y}_r$	17.35	9.00
good quality	$S_{\bar{y}}$	3.60	3.69
video-tape	Y_a	16.69	9.25
	F_a	10.56	6.67
	$p <$	.002	.025

[a] N is the sample size.

$\bar{X}_1$ is the mean score on the Wide Range Vocabulary Test, V-3

$\bar{X}_2$ is the mean score on the Necessary Arithmetic Operations, R-4

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

$S_{\bar{y}}$ is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for X_1 and X_2.

F_a is the F ratio adjusted by an analysis of covariance.

TABLE 10

Analysis of Mid-test by Lecturer and by Items Measuring the Five Sections of the Yugoslavia Article

Lecturer	Statistic[a]	Items					
		Section 1	Section 2	Section 3	Section 4	Section 5	Combined
13Y	N	59	59	59	59	59	59
	$\bar{Y}_r$	13.14	13.93	11.33	12.86	13.22	13.07
	S_y	7.98	7.38	7.23	8.39	6.88	5.57
18Y	N	48	48	48	48	48	48
	$\bar{Y}_r$	15.19	16.02	11.81	15.06	11.91	14.17
	S_y	8.56	6.43	6.71	9.10	5.54	4.68
25Y	N	58	58	58	58	58	58
	$\bar{Y}_r$	11/58	11.23	12.02	11.37	11.95	11.81
	S_y	840	6.34	8.38	7.45	6.57	5.41
31Y	N	59	59	59	59	59	59
	$\bar{Y}_r$	13.42	12.78	10.03	14.14	14.27	13.00
	S_y	6.43	6.51	7.17	7.27	6.26	4.16

Note - All scores are normalized to a maximum raw score of 30.

[a] N is the size of the sample.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_y is the mean number of items correct on the criterion tests adjusted with an analysis of covariance for X_1 and X_2.

TABLE 11

Analysis of Post-test by Lecturer and by Items Measuring the Five Sections of the Yugoslavia Article

Lecturer	Statistic[a]	Items					
		Section 1	Section 2	Section 3	Section 4	Section 5	Combined
13Y	N	103	103	103	103	103	103
	$\bar{Y}_r$	16.18	14.81	12.88	14.32	14.12	14.56
	S_y	8.20	6.87	6.71	8.84	6.62	5.35
18Y	N	62	62	62	62	62	62
	$\bar{Y}_r$	14.46	16.97	12.98	14.07	11.85	14.39
	S_y	7.21	5.28	7.05	7.25	7.27	4.38
25Y	N	98	98	98	98	98	98
	$\bar{Y}_r$	12.98	14.74	10.69	14.76	12.47	13.38
	S_y	8.34	6.47	6.40	7.84	6.41	4.86
31Y	N	80	80	80	80	80	80
	$\bar{Y}_r$	13.82	15.399	12.58	15.47	14.46	14.48
	S_y	8.17	6.54	7.04	6.80	6.71	5.08

Note: All scores are normalized to a maximum raw score of 30.

[a] N is the size of the sample.

$\bar{Y}_r$ is the mean number of items correct on the criterion tests.

S_7 is the mean number of items correct on the criterion tests adjusted with analysis of covariance for X_1 and X_2.

Bureau No 5-0252

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

(TOP)

	ERIC ACCESSION NO.				
	CLEARINGHOUSE ACCESSION NUMBER	RESUME DATE	P.A.	T.A.	IS DOCUMENT COPYRIGHTED? YES ☐ N
001		3 — — 68			ERIC REPRODUCTION RELEASE? YES ☒ N

100 TITLE

101 Validation of a Criterion of Lecture Effectiveness
102
103

200 PERSONAL AUTHOR(S)

W.P. Gorth; D.W. Allen; T.W. Stroud; L.W. Popejoy

300 INSTITUTION (SOURCE) — SOURCE CO

Stanford Center for Research and Development in Teaching

310 REPORT/SERIES NO.

320 OTHER SOURCE — SOURCE CO

330 OTHER REPORT NO.

340 OTHER SOURCE — SOURCE CO

350 OTHER REPORT NO.

400 PUB'L. DATE 4 — — 68 CONTRACT/GRANT NUMBER OE-6-10-078

500 PAGINATION, ETC.
501

27 pages

600 RETRIEVAL TERMS

lecture
601 effectiveness
602 explaining
603 videotape recording
604
605
606

607 IDENTIFIERS

800 ABSTRACT

This paper deals with several problems of research on the teacher's lecture effectiveness. In previous studies the criterion of effectiveness has been a measure of student achievement administered after presentation of live lectures. Validation of this criterion has been attempted through postdictions of mean scores on these measures from ratings of videotapes of the live lectures. The present paper discusses certain problems inherent in this procedure and investigates alternative methods.

Specifically the paper considers two basic questions: (1) whether live and videotape recordings of lectures produce the same relative mean achievement scores and (2) whether the quality of videotape recordings affects student's scores on the criterion measure. Several additional questions concerned with effects of student's ability, age of students, and the nature of the criterion itself are also investigated.